The Moral Status of AGI-enabled Robots: A Functionality-Based Analysis

Mubarak Hussain

Abstract: For a long time, researchers of Artificial Intelligence (AI) and futurists have hypothesized that the developed Artificial General Intelligence (AGI) systems can execute intellectual and behavioral tasks similar to human beings. However, there are two possible concerns regarding the emergence of AGI systems and their moral status, namely: 1) is it possible to grant moral status to the AGI-enabled robots similar to humans? 2) if it is (im)possible, then under what conditions do such robots (fail to) achieve moral status similar to humans? To examine the possibilities, the present study puts forward a functionality argument, which claims that if a human being and an AGI-enabled robot have similar functionality, but different creative processes, they may have similar moral status. Furthermore, the functionality argument asserts that an entity's (a human being or an AGI-enabled robot) creation/production from carbon or silicon or its brain's utilization of neurotransmitters or semiconductors does not carry any significance. Rather, if both entities have similar functionality, they may have similar moral status, which implies that the AGI-enabled robot may achieve human-like moral status if it performs human-like functions.

Keywords: Ethics of AI, Artificial Intelligence (AI), Artificial General Intelligence (AGI), functionalism, functionality argument, moral status.

Introduction

Suppose you and a robot work on a project together in the future. The robot is enabled with Artificial General Intelligence (AGI), which may perform intellectual and behavioral tasks similar to yours. Let us suppose the robot can think, ride, drive, cook, write a research paper, play football, and so forth, similarly to you. Despite all these functional similarities, we must admit that you are a biological entity born out of natural reproduction. However, the robot is a non-biological entity born out of programming. Both of you are functionally similar, but the creation process is different. Here, a question may arise whether it is right to lie to or mistreat the robot. Or, if one of your friends comes to the office and starts mistreating or misbehaving with the robot, will you stop your friend and ask him to be polite with the robot? Since you know that the robot is functionally similar to you (or your friend). To be precise, the robot can think similarly to you (and your friend) and understand the mistreatment or misbehaviour. Or will you not react to your friend's behavior because biologically (or physically) you both are not similar to the robot? In such a situation, if you consider it morally wrong to mistreat the robot, what kind of moral status or rights can we confer to the robot? Is it possible to grant human-like moral status to such robots? As it is assumed

that the primary purpose of developing AI is to serve the interest of humans, the moral status of robots is becoming increasingly important. One may presume that future AI may get human-like moral status by looking at the present AI systems' moral and legal rights. In 2017 a humanoid robot called Sophia, developed by Hanson Robotics, was granted citizenship by Saudi Arabia. The European Parliament proposed to confer electronic personhood and legal rights to specific AI systems. Given these developments, there is a scope to think about granting human-like moral status to the AGI-enabled robots. This paper puts forward a functionality-based approach to examine the possibility of conferring human-like moral status to AGI-enabled robots. To look into the possibilities, this paper is divided into four sections. The first section of this paper talks about the basic understanding of the concept of intelligence and AI. Based on intelligence and AI, this paper estimates the conception of AGI. The second section discusses the concept of moral status. This section looks into the Turing triage test, a hypothetical scenario introduced by Robert Sparrow to examine the importance of future AI's moral status. The third section criticizes the intelligence and sentience arguments as the criteria for conferring moral status to intelligent systems. The fourth section discusses the theory of Functionalism. Based on Functionalism, this paper puts forward a functionality argument. The argument states that if the AGI system can have human-like functionality, it may have human-like moral status.

1.(a) Intelligence

In this section, we discuss the gradual development of the concept of intelligence in the literature of psychology. The main objective of this section is to offer a rough estimation of the idea of AGI. Before sketching, what is intelligence? We must stress that intelligence is one of the most debatable subjects in psychology. Intelligence may be described but cannot be fully defined since various psychologists give various definitions of intelligence. Roughly, intelligence is the capability to reason, think logically, imagine, learn, and apply judgment. As Sparrow states, "intelligence is generalizable; it is capable of doing these things across a wide range of problems and contexts." (2004, 204). Legg and Hutter (2007) collected 70 informal definitions of intelligence in their paper *A Collection of Definitions of Intelligence*. They deduct some common features of intelligence out of the 70 definitions; *firstly*, intelligence is a quality of an individual agent through which it interacts with its environment. *Secondly*, intelligence is an agent's ability to achieve success concerning a particular goal or target. And *thirdly*, intelligence determines the ability of an individual agent to adapt to varied goals and situations. (Legg and Hutter 2007, 9). There are several intelligence theories, such as the g-factor, primary mental abilities, the theory of multiple intelligence, and the triarchic intelligence theory. Charles Spearman (Spearman 1904; quoted in Pal, Pal and Tourani 2004, 181-182) proposes the g-factor theory of intelligence, a traditional psychological indicator of intelligence.

Spearman talks about two factors: the 'g' factor, which is general intelligence, and the 's' factor, which is specific. The 's' or specific factor refers to distinct, singular, and special activities and abilities. However, all intellectual factors have a single factor, called the *g-factor* or general intelligence, that underlines all specific abilities. For instance, people may have particular talents like playing cricket, playing the harmonium, singing, writing poems, etc. All of these specific talents fall under the *g-factor*. Psychologist Louis L. Thurstone (Thurstone 1938; quoted in Cherry 2022, 2) focuses on a different theory of intelligence based on primary mental abilities. He offers seven primary cognitive abilities instead of a single general intelligence, for instance, Associative memory, Numerical ability, Perceptual speed, Reasoning, Spatial visualization, Verbal comprehension, and Word fluency. Howard Gardner (Gardner 1983; quoted in Pal, Pal and Tourani 2004, 183-184) also mentions that we do not have just an intellectual capacity. He divides intelligence into eight specialized-intelligence components (known as the *theory of multiple intelligence), namely logical-mathematical, visual-spatial intelligence, linguistic-verbal, musical intelligence, bodily-kinesthetic intelligence, interpersonal intelligence, intrapersonal intelligence, and naturalistic intelligence.* Robert Sternberg (Sternberg 1985; quoted in Cherry 2022, 3) gives a different approach where he differentiates three aspects of intelligence (*triarchic theory of intelligence) such as componential intelligence, contextual intelligence, and experiential intelligence. Componential (analytical) intelligence is the capacity to analyze information and solve issues. This intelligence includes logic, abstract reasoning, speaking ability, and mathematical ability. Contextual intelligence, also known as practical intelligence, applies information and knowledge to real-life situations.* This kind of intelligence can adapt to a changing environment. And finally, experiential or creative intelligence is the mind's capacity to learn and adapt through experience. This kind of intelligence can come up with new ideas. As a conclusion of this review of various approaches, intelligence refers to the capacity to attain goals in an extensive range of situations. Here, the ability to learn, adapt or understand is included since, through these abilities, an agent can succeed in an extensive range of situations. The subsequent section will discuss what AI means and the commonalities and differences between intelligence and AI.

1.(b) Artificial Intelligence (AI) and Artificial General Intelligence (AGI)

As we learn that there is little agreement on the definition of intelligence, similar disagreement is visible in the context of AI. In 1955, Stanford Professor John McCarthy originated the term artificial intelligence (AI) at a conference in Dartmouth. John McCarthy defines AI as the engineering and science of developing intelligent systems (Liao 2020, 3). Artificial Intelligence (in short, AI) is the sub-domain of Computer Science used to develop software programs that enable computers to exhibit intelligent behavior (Thomas 2020, 1). Or, AI reproduces

human intelligence in systems that can think like humans and imitate their activities. Bringsjord and Govindarajulu (2022) state,

> AI is the field devoted to building artificial animals (or at least artificial creatures that – in suitable contexts – appear to be animals) and, for many, artificial persons (or at least artificial creatures that – in suitable contexts – appear to be persons). (1)

Stuart Russell and Peter Norvig define AI in four different manners, such as, a) acting like humans (acting humanly), b) thinking like humans (thinking humanly, c) thinking rationally and d) acting rationally (Russell and Norvig 2010; quoted in Liao 2020, 3). They give more interest in the (iv) option, that is, acting rationally or rational action of AI. Based on such understanding, AI can take various forms. The initial form of AI is symbolic AI or [good old-fashioned AI(GOFAI)], which dominated AI research and development from 1950 to 1980. Through logic and symbolic reasoning, symbolic AI portrays cognitive tasks like thinking, learning, and problem-solving. Such AI creates the input-output relationship using a sequence of explicitly designed if-then rules. This type of symbolic AI is based on rule engines, for instance, expert systems. It also contains knowledge graphs, which are graphical representations of information stored in databases. The fundamental disadvantage of such AI is that it is problematic to change the rules or data if encoded into an AI system. Machine learning (ML) is one more form of AI which employs an algorithm to learn from various data without being explicitly programmed. There may be three kinds of ML: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is a kind of ML in which a given data set is labeled. Liao (2020) states that in supervised learning,

> an algorithm is trained on a training data set in which the correct answers for certain data are known, and the data are labeled accordingly. This way, the algorithm can use the labeled information to learn the relationship between inputs and outputs. Once the algorithm is properly trained, it is then able to apply what it has learned to predict the correct answer in different (target) data sets. (3-4)

However, unsupervised learning doesn't label a given data set. The algorithm is capable of sorting the data on its own. In this type of learning, "through clustering, an algorithm aims to group data that are more similar to each other than data in other groups." (Liao 2020, 4) Reinforcement learning is a kind of ML where the algorithm attempts to learn through experience or trial and error. There is also advanced learning called Deep learning (DL), inspired by the structure of the human brain. According to López-Rubio (2018),

Deep learning is a kind of machine learning which happens in a certain type of artificial neural networks called deep networks. Artificial deep networks, which exhibit many similarities with biological ones, have consistently shown human-like performance in many intelligent tasks. (667)

In terms of DL, this structure is called an artificial neural network. (ANN).[1] DL employs ANN, which simulates neuron activities in the brain. The primary goal of this work is to simulate the human brain, i.e., learn concepts similarly to humans. Presently, DL is the most successful ML approach.

AI may be of three kinds: artificial narrow intelligence (ANI), artificial general intelligence (AGI), and artificial super intelligence (ASI). Here we will only discuss about ANI and AGI. In short, ASI is a hypothetical intelligence in which machines could exceed human intelligence and perform any task better than humans, and it is an outcome of AGI. The term 'narrow AI' was used by Ray Kurzweil in order to specify the development of AI systems that can execute particular intelligent behavior or tasks in particular areas. (Kurzweil 2005; quoted in Goertzel 2014, 1). Narrow AI algorithms are domain or task-specific, meaning the AI algorithms that are equivalent or superior to human intelligence are intentionally programmed only in a particular or restricted field. For instance, Deep Blue of IBM becomes the world champion in Chess after beating Gary Kasparov, the world champion. Though Deep Blue wins the chess championship, it can't even play checkers. Taylor, Kuhlmann and Stone state that narrow AI differs significantly from naturally intelligent systems, such as humans, which have a wide range of abilities (Taylor, Kuhlman and Stone 2008; quoted in Goertzel 2014, 1).

The primary objective of the AI field is to create software and hardware systems that can have general intelligence. Or thinking machines that are similar to or even more than human intelligence. In recent decades wider communities of AI researchers focused increasingly on the primary goal of the AI field. The concept of AGI emerges as opposite to ANI or 'narrow AI.' (Goertzel 2014, 1). We will get a better understanding of AGI by looking at the following features of general intelligence that the AGI community agreed upon. Firstly, general intelligence refers to the capacity to attain various goals and perform multiple activities in several circumstances. Secondly, a generally intelligent system has to be capable of dealing with difficulties and situations that its developers do not predict. Thirdly, a generally intelligent system has to be capable of generalizing its acquired knowledge in order to move it from one context or problem to another context or problem (Goertzel 2014, 1-3). Goertzel and Pennachin (2007) state that

[1] The model of ANN is based on a neural network. Approximately one hundred billion neurons exist in the human brain. Every neuron is united to 1000 (approximately) neurons through synapses, giving the brain about one hundred trillion connections. An artificial neural network (ANN) is composed of artificial neurons that are simpler relative to natural ones. In pattern recognition, neural networks are efficient. For instance, it is not necessary to program the criteria used by humans to recognize a cow in an image if anyone wants to teach a neural network to do the same. In the human case, there is no problem distinguishing between two animals, i.e., cows and goats. A human may explain the distinctions to some extent; nevertheless, very few (probably no one) can provide a thorough list of all criteria employed in the identification. This is an example of tacit knowledge learned through examples and counterexamples. A similar type of learning process is employed in neural networks.

these characteristics depend on a certain level of human intelligence and hypothesize that

> the multiple specializations nature of human intelligence will be shared by any AGI systems operating with similarly limited resources, but as with much else regarding AGI, only time will tell. (7)

However, later Goertzel states that AGI is reasonably an abstract concept that is not inherently linked to any specific human features. Some qualities of human general intelligence might be universal to all-powerful AGIs, but considering our limited understanding of general intelligence, it is unclear what these might be (Goertzel 2014, 6).

We need to keep in mind that natural intelligence and artificial intelligence are different. Certainly, both may share some standard features, but they are different in many cases. In 1976 Joseph Weizenbaum, in his book, *Computer Power and Human Reason* (Weizenbaum 1976; quoted in Fjelland 2020, 2) distinguished between human reason and computer power. According to Joseph, human reason and computers are fundamentally diverse. Computer power has the capacity to employ algorithms at an incredible speed. However, human reason is based on Aristotelean prudence and wisdom. Prudence is the capacity to take or make the correct decision in specific circumstances, whereas wisdom is the capacity to grasp the entire picture. Human reason is not algorithmic. Thus, it is not possible to develop computer power into human reason. Similarly, human reason cannot substitute computer power. Roger Penrose (Penrose 1989, 1994; quoted in Fjelland 2020, 2) also mentions that our (human) thinking is not algorithmic. Certainly, the field of AGI is a new field of study on a relatively early stage of development. Based on the features discussed above, we can estimate AGI as the subfield of AI, which has the capacity to resolve general problems and handle problems and circumstances by itself. It also has the ability to acquire knowledge from other intelligent beings and their environment, like humans. As I already mentioned, AGI presently does not exist; however, the development of AGI has become a heated topic in the media and academic circles. As AI technology is evolving extraordinarily, there is a chance to develop AGI systems in the future. The future existence of AGI brings questions, such as: can the AGI achieve moral status? if so, what kind of moral status does such a system deserve? is it possible to grant it with human-like moral status or not? if we grant the AGI human-like moral status, what is the ground, or under what condition, may it be granted human-like moral status? This paper aims to give light on this kind of related questions. To do so, it is helpful to have an idea of what moral status represents.

2. Moral Status

In his paper called *The Turing Triage Test* Robert Sparrow (2004, 206-207) proposes a hypothetical situation where he talks about the moral status of future AI. The situation goes like this; you are a top medical officer in a hospital. The

hospital employs a powerful AI system for diagnosing diseases. The AI system can learn, reason, and make decisions independently. It can also have conversations with the doctors of the hospitals regarding the patients. Furthermore, the system can pass the Turing test[2] since people cannot identify whether it is a human or an intelligent machine when communicating with doctors at other hospitals over the phone (or with hospital employees or patients) via intercommunication. The hospital also has an ICU (intensive care unit) facility where six patients may benefit of the life support system. Now imagine the electricity service of the hospital is shut down due to a catastrophe. It may take some time to restore the hospital's electricity service. Then, there are only two patients in the ICU. Though the hospital has a power backup system, it is also seriously affected due to catastrophic events. The technician informs you that the available power will end very soon. However, with the available backed-up power, only one patient can survive on the full life support system. At that moment, you have to decide which patient should get continuous life support. If you don't make any decision, both patients may die soon. In the meantime, the powerful (or sophisticated) AI system also doesn't have much battery, and it might shut down at any time. At the moment, you face a 'triage' situation because you have to decide which entity (either a human patient or a powerful AI system) should get the resources. You have to make an immediate choice between the human patients and the AI system: if you want to save the patients' lives, you must switch off the AI; if you want to keep AI system 'alive', you must switch off the life support system. Switching off the AI system leads to fusing its circuit, and, as a result, it will never operate again. But if you don't make any decision immediately, both the patients and the AI system will die. The AI system signals you to plug it in in order to survive. So, in that situation, if you think it is good or reasonable to save a powerful AI system over human life, then one may say that such a powerful AI system has moral status.

However, questions may arise in this regard. Gibert and Martin argue that such a life-or-death situation shown by the hypothetical test doesn't entail the whole picture of moral status (Gibert and Martin 2022, 320). Maybe this kind of situation is helpful for moral relevance. According to them, Francis Kamm states that "X has moral status=because X counts morally in its own right, it is permissible/impermissible to do things to it for its own sake." (Kamm 2007, 227-236; quoted in Bostrom and Yudkowsky 2014, 321). He indicates that entities have moral status when they have rights and are valued for their own sake, as well as when they give us reason to maintain their existence in their own rights. An example might clarify this claim. Suppose A and B are two non-living entities, i.e., A is a gold ring, and B is a diamond ring of your marriage ceremony. Suppose fire catches at your home; you want to prevent these entities from burning, because you think it is good to save them from burning (you want to keep them), but you

[2] Oppy and Dowe (2021) state, "the phrase the Turing test is most properly used to refer to proposal made by Turing (1950) as a way of dealing with the question whether machines can think." (1)

don't think it would be good for the gold and diamond ring to continue their existence. In such cases, we may say that such entities don't have moral status but are morally significant. There may be moral reasons not to destroy or burn the diamond ring, but that doesn't suggest the diamond ring has moral status. On the other hand, we may say that a cat has some degree of moral status because it is good for a cat to continue its existence for its own sake. The sophisticated AI system described in the 'Turing Triage Test' may have moral status if one decides to preserve it for its own sake. However, the moral reason behind saving the AI systems is not similar to the moral reason behind saving entities like the gold ring and the diamond ring of your marriage ceremony. Bostrom and Yudkowsky (2014) state,

> if someone (or something) has moral status, then it is commonly agreed that the particular being has legitimate interests, that one should consider her well-being in one's decisions, and that we accept some strict moral constraints in how we treat that being; for example, the being should not be murdered or robbed, nor should anything to be done to her property without the being's consent. (321)

Thus, we may say that humans and animals have moral status. However, it is crucial to mention that an entity having a moral right or status is not identical to being a moral agent. For instance, human infants, persons with mental disabilities, and animals have some degree of (or partial) moral status, but they are not moral agents. Kantian ethics states that only adult human beings with sophisticated cognitive capacity have full moral status, but others don't have full moral status. According to Immanuel Kant, "autonomy, the capacity to set ends via practical reasoning, must be respected and grounds the dignity of all rational beings." (Kant 1785, 434, 436; quoted in Jaworska and Tannenbaum 2021, 9). Only adult humans possess such capacities and have full moral status. Human infants, persons with mental disabilities, and animals are moral patients and have partial moral status. A moral patient can be morally wronged, but they can't be morally responsible for their wrong actions. However, moral patiency comes before moral agency[3] in the human case. For instance, a human infant is a moral patient, but they have potential to be moral agents. Moreover, each moral agent is also a moral patient. The 'Turing triage test' may not play a significant role in determining future AI systems' moral status because it is possible to have a more substantial reason to save an entity that doesn't have moral status than an entity

[3] Navari (2003) states, "there are two ways in which collectives may be considered subjects of moral concern, or have moral standing. One is as moral agents. Moral agents are characterized by the possession of autonomy, rationality, and choice, as well as by the ability to take responsibility for their actions. The other is as moral patients. Unlike moral agents, moral patients may not be autonomous, they may not have reasoning capacity, nor are they necessarily in a position to make moral choices. They are entities whose chief characteristic is not that they have duties, but rather they are those to whom duties may be owed. Rather than duties, they may have rights. In any event, they have moral standing, even if they lack the usual criteria for moral agency." (1)

with moral status. For instance, one may protect their wedding ring over a plant or an animal. Similarly, if there is only one choice, one may have more substantial reasons to rescue or save a child instead of an older person while drowning. However, it doesn't mean that the older person lacks moral status. Indeed, the older person has moral status, too.

There is a difference between AI systems and other artifacts. The major difference between AI systems and other artifacts is that AIs are intelligent entities, but other artifacts are not. But if we compare an AGI system with a human being, we will find out that both AGI and humans may be intellectually similar. Though present AIs are domain-specific, general AI may develop more or less similar to human intelligence in the future. Then the question may arise if the AGI systems are more or less similar to humans, should such systems get moral status similar to humans? Or what kind of moral status should be given to the intelligent system? Or how much moral concern or regard should we have for them? To look into such queries, we need to put forward the functionality argument, which argues that the AGI system may achieve human-like moral status if it may have human-like functionality. Before proceeding to the main argument, i.e., the functionality argument, let us understand why arguments from sentience and intelligence may not be the criteria for granting moral status to the AGI system.

3. The Arguments from Intelligence and Sentience

Undoubtedly, *intelligence* is morally crucial in many cases. For instance, we apply ethical principles to personal goals, values, actions, and moral reasoning. However, intelligence may not be the only criterion for grounding moral status. In general, in the case of humans, intelligence is not needed to confer moral status. Because it is generally agreed that human infants or people with mental disabilities can be wronged even if the cognitive ability of such beings is not the same as any typical adult human. Peter Singer (Singer 1993; quoted in Gibert and Martin 2022, 324) denies intelligence as a measure of moral status. He describes an intelligence-based slave society; in a slave society, a person with a lower IQ is a slave to a higher IQ holder. Based on this observation, he maintains that such societies are unfair and consider intelligence as arbitrary as race or gender. Singer (1993) goes further and argues,

> intelligence has nothing to do with many important interests that humans have, like the interest in avoiding pain, in satisfying basic needs for food and shelter, to love and care for any children one may have, to enjoy friendly and loving relations with others and to be free to pursue one's projects without unnecessary interference from others. (23)

Furthermore, if intelligence grounds moral status, then the strong AI deserves higher moral status than human beings since, hypothetically, strong AI's intelligence would be higher than that of human beings. Therefore, intelligence may not be a measure for granting moral status to the AGI systems.

Similarly, the sentience argument is not a requirement for granting moral status to AGI systems. Roughly speaking, being sentient is similar to being conscious and is common in the domain of animals. As Gibert and Martin state, "sentience is the ability to have subjective experience, which includes perceiving and experiencing." (2022, 326) Low et al. (2012) state in the *Cambridge Declaration on Consciousness*,

> the weight of evidence indicates that humans are not unique in processing the neurological substrates that generate consciousness. Nonhuman animals, including all mammals and birds, and many other creatures, including octopuses, also possess these neurological substrates. (2)

However, some animals cannot process the capacity of sentience (mussels, oysters, sea sponges, etc.). These entities lack sentience because they do not possess complex nervous systems available to vertebrates. Even plants can show some intelligent behavior, but they lack sentience. According to Sentientism, (Gibert and Martin 2022, 327) since sentient beings possess subjective experiences of the world, such beings can be affected by positively or negatively. For example, a sentient being like a human may not want to suffer, stay alive, or be free. Because of these interests, we should behave towards them in such a way that it doesn't violate their rights. Therefore, they should get moral status.

However, Deep environmental ethicist Richard Sylvan criticizes anthropocentrism and Sentientism based on moral status in his paper *Is There a Need for a New, an Environmental, Ethic?* (Routley 1973, 205-210). There is a crucial debate in environmental ethics about what kinds of beings have intrinsic value. An entity is intrinsically valuable if it is valuable in itself. On the other hand, instrumentally, an entity is valuable if it can be used to do something else. Many people state that clean water is not intrinsically important but is instrumentally important or valuable since it is required for a good life. Humans are intrinsically important, and good life is also intrinsically important for humans. Three theories in environmental ethics are based on the distinction between intrinsic and instrumental values: Shallow Green Environmentalism or Anthropocentric or Human-centric (MacKinnon and Fiala 2015, 401-402), Mid-Green Environmentalism or Sentientism (Gibert and Martin 2022, 327) and Deep Green Environmentalism or Deep Ecology (Hyde, Filippo and Zach 2021, 19). Shallow Green Environmentalism is Anthropocentric or Human-centric as it only provides intrinsic value to human beings. However, nonhumans are also valuable if they are useful to the humans. Shallow-Green Environmentalism states that there is nothing intrinsically wrong with destroying species, felling a forest, and torturing an animal because one finds it fun to cut down trees or torture animals. However, these practices are instrumentally wrong. It means cutting down the forest is wrong because other people enjoy walking in the woods or the forest to help prevent landslides. Furthermore, torturing an animal is bad because it has a brutalizing effect on the person committing the torture.

However, in Mid-Green Environmentalism, intrinsic value or moral status is extended to all sentient creatures. The most common form of this approach is Sentientism. As Gibert and Martin state, "sentientism extends the set of entities with a moral status to many nonhuman animals, but excludes plants and ecosystems." (2022, 327)

Animals, i.e., dogs, are worthy of moral consideration in themselves. However, what about felling a forest? Sentientism states that cutting down or destroying a forest is not intrinsically wrong. On the other hand, Deep Green Environmentalism argues that all living things, all ecosystems, natural wilderness, and the earth have intrinsic values. DGE argues that it is intrinsically wrong to fell a forest. Even if nobody cares about it and no animal lives in it. Richard Sylvan (Routley 1973, 205-210) presents a thought experiment called *the Last Man on Earth* to motivate more deep ethics. This thought experiment tries to push our intuition in a different direction. Sylvan does not accept that trees have feelings or anything similar to that. Nevertheless, he still believes that trees have intrinsic value.

The Last Man on Earth Argument:

(a)The First Form: Let us imagine that all humans are dead except a man. Let us consider him as the last man on earth. The last man wants to kill or destroy everything just before his death. He tries to destroy or kill every living plant, animal, bacteria, and so forth on the planet through powerful technology. If we look at the traditional views, the last man does nothing wrong. However, many people may have objections to what the last man does. This indicates that we need a new ethics where nonhuman nature is treated as valuable and independent of our interests. This new ethics implies that nature also has intrinsic value. However, some may argue that if trees are only instrumentally valuable, then it is not counterintuitive to state that the last man is wrong. People may judge the last man for his doing while not accepting or adopting a deep green ethics. One could argue that the problem is simply that he kills so many sentient animals. We can further reshape or even give more strength to our thought experiment.

(b)The Second Form: Let us imagine that all humans are dead except a man. Let us consider him as the last man on earth. The last man wants to kill the rest of the living things before his death. This time he destroys the entire planet, including bacteria, fungi, etc. Here, does the last man do something wrong? If so, then we need a deep green ethics. In this case, we reconcile anthropocentrism and Sentientism with our intuition and state that the last man is wrong because he kills all potential future intelligent or sentient creatures. Both theories argue that only intelligent or sentient beings are intrinsically valuable. On the other hand, non-intelligent or non-sentient life is instrumentally valuable since one-day such entities might help intelligent or sentient life evolve in the world. The last man destroys this possibility. It means the last man is wrong in preventing the potential

future intelligent or sentient creatures from evolving, but not wrong in destroying all the nonhuman creatures. Our thought experiment can be modified further.

(c)The Third Form: Let us imagine that all humans are dead except a man. Let us consider him the last man on earth. In this form, it is acknowledged that the sun will die in one million years, and sentient creatures have no scope to evolve in the future. The last man decides to kill all the creatures before his death. In this form, we are only considering non-sentient entities. If you believe that the last man did anything wrong in this case, it will be tough for you to resist a deeper green ethics. Sylvan distinguishes two values to solve the problems: Sole value assumptions and Greater value assumptions (Hyde, Filippo and Zach 2021, 17-18). According to the Sole value assumption, only humans and human projects have intrinsic value. However, the Greater value assumption is different. Although the Greater value assumption considers nonhuman things to have intrinsic value, human value always outweighs this value. It implies that whenever there is a clash between nonhuman ideals and human goals, the latter must always take priority. Richard Sylvan and other Deep environmentalists rejected both Sole and Great value assumptions (Hyde, Filippo and Zach 2021, 19). Still, Greater value assumptions of something are intuitive, and the last man argument cannot push our intuitions in a different way. The Greater value assumption may not apply when you have to choose between saving a human, i.e., Hitler, and a nonhuman animal, i.e., a pet cat. Many of us would say it would be good to save a nonhuman being, i.e., a pet cat, over a human-like Hitler. In such scenarios, we could argue that Hitler violated other people's rights and cruelly killed a particular group of people. So, he deserves to die, since people have the responsibility not to violate other people's rights and so on. In this case, the Greater value assumption is not applicable. We can modify the last man's argument further.

(d)The Fourth Form: Let us imagine that all humans are dead except a man. Let us consider him the last man on earth. In this form, it is acknowledged that the sun will die in one million years, and there is no scope for sentient creatures to evolve in the future. Further, imagine that the only remaining non-sentient entities are kept in an underground laboratory. Earth's surface is entirely destroyed due to climate change or environmental-related issues. The last man also lives in the underground laboratory and dies within a week, no matter what. He has two choices: either release some organisms to the surface to repopulate the Earth and die tomorrow, or eat them and die next week. Here, the conflict between nonhuman values and human needs is clearly visible. In this case, many people would have the intuition that the last man should give up that week and release nonhuman animals. Suppose the last man releases the nonhuman beings. In that case, the Greater value assumption (whenever there is a clash between nonhuman ideals and human goals, the latter must always take priority) must be wrong, because it contradicts what it initially states. This thought experiment gives some sense of why people might be inclined to deeper environmental ethics and why being intelligent and sentient may not be a criterion for having moral status.

Now problematic questions may arise in the context of the possibility of developing sentient AI. Some people predict that sentient AI is possible through a hypothetical technology called brain emulation or uploading[4]. Even science fiction movies such as 'Transcendence' (2014) also show the possibility of sentient AI. Currently, there is an AI system called Shelly, a robot tortoise designed to mimic pain. An AI system like Shelly can react like a sentient being. However, it can only mimic without authentic feeling. Apart from such issues, arguments from sentience can be criticized on the grounds stated by the *problem of other minds*. Nagel (1987) asks, "can you really know about the conscious life in this world beyond the fact that you have a conscious mind?" (26-27) The philosophical problem of other minds can be condensed in the following terms: I am aware that I have a mind and have mental status, i.e., feelings, sensations, and thoughts. I have direct access to know such mental states, or I have direct awareness about the mental states. However, how do I know other people also have a mind and such mental awareness in their mind? I can only observe others' behavior, but how can I know others also have minds? We cannot directly access other minds because the knowledge of other minds is always indirect. Now let us take an example to understand whether sentient AI is possible. We do not have any access to know how exactly a cat experiences; similarly, though we know what precisely an AI is, we do not know how exactly an AI experiences certain things. We can understand better an AI system than a cat. A cat might be a more complex entity. Cat experiences and human experiences may have very slight similarities, and thus, in general, when a cat seems to be in pain, the humans can, at least in a minimal sense, understand what it is to have pain. There are some biological similarities between a human and a cat, but these similarities do not apply to the human being and the AI system. If an AI looks as if experiencing pain, we certainly cannot tell if it is actually in pain. Regardless of the issue of other minds in the context of animals, we can decide with certainty if a cat is really in pain. Therefore, the sentience argument may not be a criterion for conferring moral status to the AGI systems. Moreover, there is no substantial evidence that a sentient AI will exist. In the next section, we look into the argument from functionality for conferring moral status to the AGI systems. Before going into the functionality argument, let us glimpse the theory of Functionalism.

4. (a) Functionalism

Functionalism is one of the most famous theories in the philosophy of mind, which deals with the nature of mental states. As Polger (2022) states,

[4] Bostrom and Yudkowsky (2014) state, "uploading refers to a hypothetical future technology that enable a human or human or other animal intellect to be transferred from its original implementation in an organic brain onto a digital computer." (325)

> Functionalism is a theory about the nature of mental states. According to functionalism, mental states are identified by what they do rather than by what they are made of. (1)

The main idea of Functionalism can be explained by taking the example of an ATM and a diamond. ATMs could be made out of plastic, metal, or any other material, but the job or function of ATMs is to help withdraw money. Because of this reason, an ATM can be considered as having a functional organization. Similarly, the functionality of a caretaker is to assist with personal care, i.e., bathing and grooming, preparing meals, shopping, housekeeping, and so on. A caretaker may be anyone; it may be me, it may be another person, or it may be X, Y, and Z, or even it may be an intelligent robot. Whoever may be the caretaker, if anyone can perform the expected job mentioned above, then he/she is fulfilling the functionality of a caretaker.

However, we can't keep a diamond in the same category since diamonds are particular physical objects that consist of molecular lattice structures and carbon crystals. Without such specific materials, a diamond can't be made out. Functionalism states that mental states are similar to ATMs but not to diamonds. Some things may be created or proven to exist based on how they relate to other things and their features. A key may be physical stuff with a particular composition, but being physical stuff with a particular composition doesn't matter much. The main thing that matters for a key is whether it can perform a specific action, such as opening a lock. Similarly, a lock is a kind of material that exists in connection to keys. There may be various types of keys, such as metal, wooden, plastic, and digital keys. Functionalism states that what makes something a mental state is not what it is made of but what it does. In the case of a key, the material composition of a key doesn't make it a key but rather what it does or can do. The actions a key performs or is expected to perform are referred to as its functions. Opening a lock is a function of a key; that's why they are functional entities or a functional kind.

The original idea of Functionalism arises from the comparison of minds with computers. But the comparison between minds and computers is just an analogy. Functionalism comes as a substitute for behaviorism and the identity theory of mind. As already mentioned, what makes something a mental state is not what it is made of, but what it does, and mental states are more like ATMs than diamonds. This statement distinguished Functionalism from Descartes's mind-body dualism. As mentioned by Polger (2022, 2), Descartes states that the mind is made of a specific substance called res cogitans or thinking substance. Functionalism is different from behaviorism which holds that to be in a mental state is merely to exhibit certain kinds of behavior. As against behaviorism, it argues that mental states must be inner states with functional-causal roles. It is also different from the mind-brain identity theory. According to the mind-brain identity theory, mental states are specific kinds of biological states of the brain (Polger 2022, 2). Thus, mental states are made up of some brain stuff. Mind-brain

identity theory states that mental states are more similar to diamonds than ATMs. Opposing mind-brain identity theory, Functionalism brings a more liberal approach through which any entity/being/system, for instance, computer program, souls, extra-terrestrials, etc., can pass the *Turing Test* is eligible as an intelligent being (Gokel 2013, 13).

According to Shagrir (2005) Hilary Putnam's computational functionalism is the theory of mind which holds that:

> mental states and events-pains, beliefs, desires, thoughts and so forth-are computational states of the brain and so are defined in terms of computational parameters plus relations to biologically characterized inputs and outputs. (1-3)

The nature of the brain is not dependent on its physical structure. He further mentions, "we could be made of Swiss cheese, and it wouldn't matter." (2) The only thing that matters is functional organization. A brain could be made out of different materials, made out of metal or wood; what is important is how mental states are causally related to one another, for instance, sensory inputs and motor outputs. Some things, such as trees, stones, and hearts, don't have minds. But why do such things not have minds? The answer to this question is that such things don't have the right functional organization. This implies that other thinking entities may be made of Swiss cheese along with the right (or suitable or appropriate) functional organization. Putnam considers mental states as functional states (Putnam 1967a, 1967b; quoted in Shagrir 2005, 3). According to Putnam, it is helpful to consider minds as machines of a specific sort. Functionalism was advanced as a reply to the mind-brain identity theory. As Smart (1959) states, "sensations are brain processes" (144). It indicates that mental states and brain states will have a one-to-one relationship if we consider mental states to be brain states. However, functionalists argue that mental and brain states are not identical. Were they identical, everything with sensation S would have brain state B, and everything with brain state B would have sensation S. Functionalism denies the mind-brain identity theory. Mammals, reptiles, and mollusks can have the ability to feel pain, but such entities do not have brains similar to humans. Thus, there is no one-to-one relationship between sensations and brain processes. Though mammals, reptiles, and mollusks have different brains, they still perform the same action or function similarly. Therefore, it is not necessary to have a one-to-one relation between mental states and brain states. Functionalism holds that minds are mechanisms, and there are multiple ways to construct such mechanisms (Polger 2022, 5).

(b) The Functionality Argument

Let us put forward a functionality argument to look into the possibility of the moral status of AGI-enabled robots. The functionality argument states that if two entities, such as a human being and an AGI-enabled robot, have similar functionality, but the creation process of both entities is different, then they may

have similar moral status. As already discussed, mental states are recognized by what they execute (or do) rather than by how they are formed. The argument from functionality states that the creation process of an entity does not bear the moral status of whether an entity is produced from carbon or silicon. It also doesn't matter whether the brains of such entities use semiconductors or neurotransmitters. If the AGI system can have human-like functionality, it may be granted human-like moral status. In the past, the moral status of a person was dependent on their caste, gene, or bloodline. Alternatively, there could be other criteria, i.e., *intelligence, sentience*, and so forth, through which people confer moral status to an entity. However, the argument from functionality is against such bases of moral status. It argues that causal factors, for instance, in vitro fertilization (IVF), assisted delivery, family planning, gamete selection techniques, etc., do not affect the moral status of a baby. Using such techniques, one may have deliberate choice and design in creating human beings. Human babies born through these technologies also have equal moral status, similar to normal human babies. Even those who oppose human reproductive cloning on religious grounds agree that cloned and natural babies should have equal moral status. Bostrom and Yudkowsky (2014, 322) mention that denying the argument from functionality would be similar to a situation where people support racism, believing that they are superior to other communities or ethnic groups. On the other hand, technology may create a fetus without a brain. In that case, a baby without a brain or an anencephalic baby does not have equal moral status, similar to a normal baby. In the case of an anencephalic baby, the function is disabled. One has a brain, and the other does not have a brain. In this case, the argument from functionality may not apply. Here, one may argue in the following manner: if we suppose that the concept of functionality is fundamental in deciding the moral status of something, then an anencephalic baby will not have any moral status. However, it seems counterintuitive to say that we can render any treatment to an anencephalic baby since it does not have a brain, because of which there is a difference in functionality. Responding to such queries, one may argue that an anencephalic baby may not have a similar moral status as a normal baby. However, such a baby will have some degree of moral status**.** It is not argued that an anencephalic baby will not have any moral status and that we can render any treatment to such a baby. As mentioned earlier, an entity having a moral status is not equal to being a moral agent. For instance, human infants, persons with mental disabilities, and animals have some degree of moral right or status. However, they are not considered moral agents but are moral patients. An anencephalic baby is a moral patient; a human baby is also a moral patient. Nevertheless, these two babies are different; an anencephalic baby does not have the potential to develop a cognitive capacity (most anencephalic babies die before birth; if born, they die within a few hours, days, or weeks). On the other hand, normal babies have the potential to develop a cognitive capacity in the future.

However, authors like Alison Davis do not find any dissimilarity between these two babies. Davis (1998) writes,

> in my view, there can be no sound differentiation between the two, and that being so, I believe individual rights begin when individual lives begin-at conception- and should be protected from then on. Transplants from those other than anencephalic are subject to very strict rules, and the donor must have consented and/or be physically dead. I can see no reason why anencephalic should be treated any differently. They are not physically dead when used as donors, and are in any case incapable of consenting. (151)

Although anencephalic babies do not have cognitive capacity (or are unable to develop cognitive ability), they have specific functionality, i.e., they have a brain stem that regulates respiration and reflex movements. In such cases, the anencephalic baby may be granted some degrees of moral status in virtue of functionality. As we discussed already, AGI system does not exist currently. However, a rough estimation of the feature of an AGI system could be given. The rough estimation entails that functionally an AGI system may be similar to human beings. The argument from functionality states that though AGI would be different from humans, it may still exhibit similar intellect and behavior to humans. It does not matter whether an AI system is developed or born out of programming or runs on a computer instead of in a brain. The AGI system may deserve human-like moral status if it can have human-like functionality.

Conclusion

The main objective of this paper was to discuss the future AI's (or AGI's) moral status. Regardless of not having the certainty concerning the development of such systems, there is already a heated debate in the media and the academic circle. Gradually AI systems have become more powerful and adaptable and can execute different cognitive tasks. For instance, identifying objects from videos and images, translating various languages, stock trading, driving automobiles, drawing their encryption language, identifying cancer in tissues, and so on.

Apart from these advancements in AI technology, three milestones in AI technology have gained public attention. Based on such milestones, one may assume that AGI is knocking on our door; however, the development of AGI is still in an early stage. The first milestone of AI technology is the Deep Blue of IBM. The Chess algorithm Deep Blue won the Chess championship after beating the world champion, Garry Kasparov, in 1997. Even though Deep Blue performed extraordinarily well, it is still a narrow AI since it could only play Chess but could not even play checkers.

The second milestone is IBM's AI program, Watson**.** IBM developed a computer program called Watson to participate in a quiz show called Jeopardy. Contestants on Jeopardy were given the possible answers and were then expected to find the correct answer. Jeopardy is more complex than Chess since it requires a wider variety of knowledge and skills. This game includes various areas of

expertise, i.e., geography, history, science, sports, and culture. Analogies and puns are also included in the game. Since its beginning in 1964, the quiz show has gained enormous popularity in the US. There were three participants in Jeopardy. If the first participant answered incorrectly, the second could answer the question. The quiz program participants needed the knowledge, speed, and capacity to limit themselves. Watson's communication was based on NLP or NLU[5]. According to IBM cloud education,

> natural language processing (NLP) strives to build machines that understand and respond to text or voice data- and respond with text or speech of their own- in much the same way humans do. (2020)

Most importantly, Watson was not enabled with internet while playing the Jeopardy game. However, Watson had access to almost 200 million pages of information. Watson defeated Ken Jennings and Brad Rutter in 2011. Jennings won 74 consecutive races in 2004 and earned $3 million in prize money. In 2005 Rutter defeated Ken and received 3 million US dollars. Further, IBM wanted to develop an AI medical super-doctor. IBM thought that if Watson could access all medical data like medical records of patients, journal papers, drug lists, etc., then Watson may give better diagnoses and treatment than human doctors. However, that could not happen in reality.

The third and latest milestone of AI technology is DeepMind's AlphaGo. Go is a board game that originated in China about 2000 years ago and is one of the most complex games, considered more challenging than Chess. The game Go is mainly played in East Asian countries, i.e., China, Japan, Mongolia, North Korea, and Taiwan. AlphaGo, which was built on an advanced search tree and deep neural networks, defeated three-time Go champion of Europe Fan Hui in 2015 and eighteen-time world champion Lee Sedol in 2016. DeepMind has gradually launched an enhanced version of AlphaGo called AlphaGo Zero. It was trained by playing against itself, beginning with purely random play. In late 2017, DeepMind again introduced an extended version of the algorithm known as AlphaZero. This algorithm also taught itself. MuZero is the most recent version of DeepMind's algorithm. It performs the game similar to AlphaZero in Go, Chess, and Shogi. MuZero also learns to master various visually complicated Atari games without being taught the rules of any of them. AlphaGo utilizes Deep reinforcement learning, which is based on the ANN model. There is a significant difference between Deep Blue and AlphaGo. Let us first compare human chess players and Deep Blue. Human chess players use intuition and calculation. Through these two

[5] "Watson Natural Language Understanding (NLU) - analyze text in unstructured data formats including HTML, webpages, social media, and more. Increase your understanding of human language by leveraging this natural language tool kit to identify concepts, keywords, categories, semantics and emotions, and to perform text classification, entity extraction, named entity recognition (NER), sentiment analysis and summarization."(https://www.ibm.com/cloud/learn/natural-language-processing#toc-natural-la-H2GEqPVg)

capacities, one can estimate a particular board position. However, Deep Blue was designed to compute many different board positions and determine the best potential positions in a specific situation. On the other hand, the scene was different in the context of the game Go. AlphaGo illustrated that algorithms could handle tacit knowledge[6]. However, Fjelland argues that AI's tacit knowledge is different from humans. The tacit knowledge of AI is limited to the idealized realm of science. The traditional AI programs' parameters were explicit and transparent. However, Deep Neural Networks are not transparent because one may not understand what parameters are used in the systems. That is why AlphaGo is considered one of the significant milestones in AI technology and shows that technology is advancing rapidly. Here one may argue that even though AlphaGo is an excellent in-game playing domain in having learned to play the game without being taught the rules, it is still a narrow AI system. We may not draw an appropriate comparison between human beings and AI systems yet. It might carry the comparison with earlier AI systems. Therefore, this advancement can't entail that AGI is very near, or that we can compare its intelligence with human intelligence. We need a more convincing set of reasons to claim that such game-playing abilities indicate such models evolving to AGI. As already mentioned, the primary goal of this paper was to investigate whether it is possible to confer human-like moral status to future AGI-enabled robots or not when they come into existence. The AGI is a hypothetical technology in AI that may be more or less intellectually similar to human beings. Therefore, we may say that if the future AI or AGI shows human-like functionality, it may have human-like moral status.

Acknowledgment

I am immensely grateful to Prof. Jolly Thomas (Department of Humanities and Social Sciences, Indian Institute of Technology Dharwad) for supervising the entire research. I thank Prof. Don Wallace Freeman Dcruz (Department of Humanities and Social Sciences, Indian Institute of Technology Delhi) for insightful remarks and recommendations for the research.

References

Baum, Seth. 2017. "A Survey of Artificial General Intelligence Projects for Ethics." *Global Catastrophic Risk Institute Working Paper* 17(1).

Bostrom, Nick, and Eliezer Yudkowsky. 2014. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith

[6] "Michael Polanyi, in his book, *Personal Knowledge* introduced the expression tacit knowledge. Most of the knowledge we apply in everyday life is tacit. In fact, we do not know which rules we apply when we perform a task. Polanyi used swimming and bicycle riding as examples. Very few swimmers know that what keeps them afloat is how they regulate their respiration: when they breathe out, they do not empty their lungs, and when they breathe in, they inflate their lungs more than normal." (Fjelland 2020, 3)

Frankish and William M. Ramsey, 316–34. Cambridge: Cambridge University Press.

Bringsjord, Selmer and Naveen Sundar Govindarajulu. 2022. "Artificial Intelligence." In *The Stanford Encyclopedia of Philosophy*. Available at: https://plato.stanford.edu/archives/fall2022/entries/artificial-intelligence/. Accessed: October 20, 2022.

Broom, Donald. 2016. "Considering Animals' Feelings: Précis of Sentience and Animal Welfare." *Animal Sentience* 5(1): 1-11.

Cherry, Kendra. 2022. "Theories of Intelligence in Psychology." Verywell mind. Available at: https://www.verywellmind.com/theories-of-intelligence-2795035. Accessed: May 7, 2023.

Davis, Alison. 1988. "The Status of Anencephalic Babies: Should Their Bodies be Used as Donor Banks?". *Journal of Medical Ethics* 14 (3): 150-153.

Everitt, Tom, Gary Lea and Marcus Hutter. 2018. "AGI Safety Literature Review." *Australian National University.* Available at: https://arxiv.org/abs/1805.01109. Accessed: April 22, 2021.

Fjelland, Ragnar. 2020. "Why General Artificial Intelligence Will Not Be Realized." *Humanities and Social Sciences Communications* 7, 10.

Gardner, Howard. 1983. *Frames of Mind: The Theory of Multiple Intelligence.* New York: Basic Books.

Gibert, Martin and Martin Dominic. 2022. "In Search of The Moral Status of AI: Why Sentience is a Strong Argument." *AI & Soc* 37: 319-330.

Goertzel, Ben, and Cassio Pennachin. 2007. *Artificial General Intelligence.* Heidelberg: Springer Berlin.

Goertzel, Ben. 2014. "Artificial General Intelligence: Concept, State of the Art, and Future Prospects." *Journal of Artificial General Intelligence* 5 (1): 1-48.

Gokel, Nazim. 2013. "Artificial Psychology, Functionalism, and Mental Representation." *Procedia – Social and Behavioral Sciences* 82: 12-18.

Gordon, John-Stuart. 2020. "What Do We Owe to Intelligent Robots?." *AI & Society* 35 (1): 209-223.

Google DeepMind. n.d. "AlphaGo." Accessed: May 15 2021. Available at: https://www.deepmind.com/research/highlighted-research/alphago. Accessed: February 20, 2022.

Hakli, Raul and Pekka Mäkelä. 2019. "Moral Responsibility of Robots and Hybrid Agents." *The Monist* 102 (2): 259-275.

Harris, Jamie and Jacy Reese Anthis. 2021. "The Moral Consideration of Artificial Entities: A Literature Review." *Science and Engineering Ethics* 27, 53.

Himma, Kenneth Einar. 2009. "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to be a Moral Agent?" *Ethics and Information Technology* 11: 19-29.

Hyde, Dominic, Filippo Casati and Zach Weber. 2021. "Richard Sylvan [Routley]." In: *The Stanford Encyclopedia of Philosophy,* edited by Edward N. Zalta.

Available at: https://plato.stanford.edu/archives/spr2021/entries/sylvan-routley/. Accessed: March 18, 2022.

IBM cloud education. 2020. "What is Natural Language Processing (NLP)?." Available at: https://www.ibm.com/cloud/learn/natural-language-processing#toc-natural-la-H2GEqPVg. Accessed: February 20, 2022.

Jaworska, Agnieszka and Julie Tannenbaum. 2021. "The Grounds of Moral Status." In: *The Stanford Encyclopedia of Philosophy,* edited by Edward N. Zalta. Available at: https://plato.stanford.edu/archives/spr2021/entries/grounds-moral-status/&gt. Accessed: June 3, 2020.

Kant, Immanuel.1785. *Groundwork of the Metaphysics of Morals*, M. Gregor (trans. and ed.). Cambridge: Cambridge University Press, 1998.

Kamm, Frances. 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm.* New Work: Oxford University Press.

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking Penguin.

Legg, Shane and Marcus Hutter. 2007. "A Collection of Definitions of Intelligence. Advances in Artificial General Intelligence: Concepts, Architectures, and Algorithms." *Frontiers in Artificial Intelligence and Applications*, 157: 17-24.

Levin, Janet. 2018. "Functionalism." In *The Stanford Encyclopedia of Philosophy,* edited by Edward N. Zalta. Available at: https://plato.stanford.edu/archives/win2021/entries/functionalism. Accessed: April 9, 2020.

Liao, S. Matthew. 2020. *Ethics of Artificial Intelligence*. New York, NY: Oxford University Press.

López-Rubio, Ezequiel. 2018. "Computational Functionalism for the Deep Learning Era." *Minds and Machines* 28: 667-688.

Low, Philip et al. 2012. *The Cambridge Declaration on Consciousness*. Publicly proclaimed in Cambridge, UK, on July 7, at the Francis Crick Memorial Conference on Consciousness in Human and non-Human Animals. Available at: https://fcmconference.org/img/CambridgeDeclarationOnConsciousness .pdf. Accessed: September 20, 2021

MacKinnon, Barbara and Andrew Fiala. 2015. *Ethics: Theory and Contemporary Issues, Eighth Edition.* CT 06902 USA: Cengage Learning.

Nagel, Thomas. 1987. *What Does It All Mean?*. New York, NY: Oxford University Press.

Navari, Cornelia. 2003. "When Agents Cannot Act: International Institutions as Moral Patients." In *Can Institutions Have Responsibilities?*, edited by Toni Erskine. Global Issues Series. London: Palgrave Macmillan.

Oppy, Graham and David Dowe. 2021. "The Turing Test." In *The Stanford Encyclopedia of Philosophy*, edited by N. Zalta. Available at: https://plato.stanford.edu/archives/win2021/entries/turing-test. Accessed: March 19, 2022.

Pal, H.R., A. Pal, and P. Tourani. 2004. "Theories of Intelligence." *Everyman's Science* XXXIX (3): 181-186.

Polger, Thomas W. 2022. "Functionalism." In *The internet Encyclopedia of Philosophy ISSN 2161-002.* Available at: https://iep.utm.edu/functism/. Accessed: March 6, 2022.

Penrose, Roger. 1989. *The Emperor's New Mind. Concerning Computers, Minds, and the Laws of Physics.* Oxford: Oxford University Press.

_____. 1994. *Shadows of the Mind. A Search for the Missing Science of Consciousness.* Oxford: Oxford University Press.

Putnam. Hilary. 1967a. "The Mental Life of Some Machines"*.* In *Intentionality, Minds and Perception*, edited by Hector-Neri Castañeda, 177-200. Detroit: Wayne State University Press.

_____. 1967b. "The Nature of Mental States" (originally published as "Psychological Predicates"). In *Art, Mind and Religion*, edited by W. H. Captain and D. D. Merrill, 37-48*,* Pittsburgh: University of Pittsburgh Press.

Risse, Mathias, and Steven Livingston. 2019. "The Future Impact of Artificial Intelligence on Humans and Human Rights." *Ethics and International Affairs* 33 (2): 141-158.

Routley, Richard. 1973. "Is There a Need for a New, an Environmental, Ethic?." In *Proceedings of the XVth World Congress of Philosophy 17th to 22 September*, Varna, Bulgaria: Sofia Press, 205-210.

Russell, Stuart J. 2010. *Artificial Intelligence: A Modern Approach, Third Edition.* Upper Saddle River, New Jersey: Pearson Education, Inc.

Scheessele, Michael. 2018. "A Framework for Grounding the Moral Status of Intelligent Machines. Artificial Intelligence." In *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 251-256.

Schwitzgebel, Eric and Mara Garza. 2015. "A Defense of the Rights of Artificial Intelligences." *Midwest Studies in Philosophy* 39 (1): 98-119.

Shagrir, Oron. 2005. "The Rise and Fall of Computational Functionalism." In *Contemporary Philosophy in Focus: Hilary Putnam*, edited by Yemima Ben-Menahem, 220–50. Cambridge: Cambridge University Press.

Singer, Peter. 1993. *Practical Ethics.* Second Edition. Cambridge: Cambridge University Press.

_____. 2009. "Speciesism and Moral Status." *Metaphilosophy* 40 (3/4): 567-581.

Steunebrink, Bas, Pei Wang and Ben Goertzel (eds.). 2016. *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, Proceedings*. Springer.

Sparrow, Robert. 2004. "The Turing Triage Test." *Ethics and Information Technology* 6: 203-213.

Smart, J. J. C. 1959. "Sensations and Brain Processes." *The Philosophical Review* 68(2): 141-156.

Sternberg, Robert J. 1985. *Beyond IQ, A Triarchic Theory of Human Intelligence.* CUP Archive.

Spearman, Charles. 1904. "General Intelligence Objective, Objectively Determined and Measured." *American Journal of Psychology* 15: 201-293.

Talbert, Matthew. 2022. "Moral Responsibility." In: *The Stanford Encyclopedia of Philosophy,* Edward N. Zalta (ed.). Available at: https://plato.stanford.edu/Archives/fall2022/entries/moral-responsibility/. Accessed: November 20, 2021.

Taylor, Matthew & Kuhlmann, Gregory & Stone, Peter. 2008. "Transfer Learning and Intelligence: An Argument and Approach." In: *Proceedings of the 2008 Conference on Artificial General Intelligence 2008: Proceedings of the first AGI conference,* 326-337.

Thomas, Richmond. 2020. "Logic and Artificial Intelligence." In *The Stanford Encyclopedia of Philosophy,* edited by Edward N. Zalta. Available at: https://plato.stanford.edu/archives/sum2020/entries/logic-ai. Accessed: May 5, 2023.

Thurstone, Louis L. 1938. *Primary Mental Abilities.* Chicago: University of Chicago Press.

Weizenbaum, Joseph. 1976. *Computer Power and Human Reason.* San Francisco: Freeman & Company.